Hits-and-misses for the evaluation and combination of forecasts

Thomas Wenzel

Department of Statistics, University of Dortmund, Vogelpothsweg 87

D-44221 Dortmund, Germany

e-mail: wenzel@statistik.uni-dortmund.de

Abstract: Error measures for the evaluation of forecasts are usually based on the size of the forecast errors. Common measures are e.g. the Mean Squared Error (MSE), the Mean Absolute Deviation (MAD) or the Mean Absolute Percentage Error (MAPE). Alternative measures for the comparison of forecasts are turning points or hits-and-misses, where an indicator loss function is used to decide, if a forecast is of high quality or not. Here, we discuss the latter to obtain reliable combined forecasts.

Key words: Combination of forecasts, hits-and-misses, turning points, fuzzy.

Acknowledgement: Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

AMS 1991 Subject classification: 62F10

1. Introduction

The basic paper dealing with the combination of forecasts from Bates and Granger (1969) discusses the calculation of optimal combination weights under the MSE-criterion which is also done in many other articles. Wenzel (1998) considers the Pitman-closeness criterion for the evaluation of forecasts. There, the Pitman-closest forecast combination is equivalent to the MSE-optimal forecast combination. Russell and Adam (1987) among other things employ the MAD. Klapper (1998) uses the ranks of prediction errors to calculate a forecast combination whereas Cicarelli (1982) describes combination methods on the basis of turning points. He counts how often a forecast had the "right" direction in the past and uses this to calculate weights. These are only a few of a number of authors who evaluate forecasts and derive combination weights under a special evaluation criterion. An overview about 28 evaluation criteria is given e.g. in Dammers (1993). Although there exist several error measures we want to discuss the not so well-known hits-and-misses criterion. We define when a forecast is said to be "good" (hit) or "bad" (miss), a criterion easily applicable in practice. Especially non-statisticians are often interested in easy calculations and dislike measures like MSE or MAPE. In practice, there is often the situation where a special decision is based on a given forecast, for instance *should we invest in stocks of a specific trade* or *which kind of clothes do we need tomorrow*? Thus, a lot of analysts would categorize the forecast in "good" or "bad", depending on the decision they made. For the first of these questions the evaluation of the rate by turning points is appropriate. Secondly, we are interested in a relatively "precise" forecast of the temperature. Thus we could consider the hits-and-misses criterion. Furthermore, we analyse a generalized hits-and-misses criterion. Here, we define not only hits and misses but also approximative hits. We discuss how to use the several approaches for the calculation of a combined forecast. Another approach is to compare the individual forecasts by statistical tests, e.g. a comparison of the error variances. Flores (1986, 1989) presents the utilization of the sign test and the Wilcoxon test for the comparison of forecasting methods in the M-competition. It is also possible to use the hits-and-misses criterion to test if there is a significant difference in the quality of the forecasts. Afterwards we consider the results for the calculation of a forecast combination.

Finally, we analyse an example given in Klapper (1998) dealing with German macro economic data to give insight how the different methods perform in practice. Furthermore, we also perform a small simulation study for the combination of two biased forecasts.

2. The hits-and-misses criterion and the combination of forecasts

The original definition of hits-and-misses for forecasts of a 0-1-variable is given in Armstrong (1985, p.353). If a forecast has the same value as the variable it is a hit, otherwise a miss. If the variable to be forecasted is continuous we have to deal with hit-intervals which is decribed in the following.

Definition 1: Let F_{T+1} be a forecast for Y_{T+1} (T+1: time index) at time T. Then F_{T+1} is a γ-hit, $\gamma \in [0,\infty)$, iff

$$F_{T+1} \in [Y_{T+1} - \gamma|Y_{T+1}|, Y_{T+1} + \gamma|Y_{T+1}|].$$

Restricting γ to the interval [0,1] we can call a forecast which is in the hit-interval a $100\cdot(1-\gamma)$-percent-hit. We can now use the hits-and-misses criterion for the combination of forecasts. Let $F_{1,T+1},\ldots,F_{n,T+1}$, $n \geq 2$, be forecasts for Y_{T+1} and let F_{it}, Y_t, $i = 1,\ldots,n$, $t = 1,\ldots,T$ be given past observations. Then we define combination weights by

$$w_{i,\gamma} := \frac{\sum_{t=T-v+1}^{T} I_{[Y_t-\gamma|Y_t|,\, Y_t+\gamma|Y_t|]}(F_{it})}{\sum_{j=1}^{n}\sum_{t=T-v+1}^{T} I_{[Y_t-\gamma|Y_t|,\, Y_t+\gamma|Y_t|]}(F_{jt})}, \quad i = 1,\ldots,n,$$

where $v \in \{1,\ldots,T\}$ is the number of past observations fixed for the calculation. The weights are well defined if there exists a forecast with at least one γ-hit in the past. This can be assured by selecting γ appropriately. For each weight we divide the number of hits of the special individual forecast by the number of hits of all forecasts. By definition $\sum_{i=1}^{n} w_{i,\gamma} = 1$, and if the individual forecasts are unbiased, the combination is also unbiased. If we again look at Definition 1, it is also possible to define hit intervals which have not Y_{T+1} as midpoint, that is $\left[Y_{T+1} - \gamma_1|Y_{T+1}|, Y_{T+1} + \gamma_2|Y_{T+1}|\right]$, $\gamma_1, \gamma_2 \geq 0$, $\gamma_1 \neq \gamma_2$. An example, where we should consider such intervals is given in Schneider, Klapper, and Wenzel (1999). For the evaluation of forecasting methods in goods management systems underestimation is punished more than overestimation, and thus we should select $\gamma_2 > \gamma_1$. Obviously, for "small" values of Y_{T+1} the hit-interval is also "smaller". Therefore, we give an alternative definition of hits-and-misses.

Defintion 2: Let F_{T+1} be a forecast for Y_{T+1} at time T. Then F_{T+1} is a c-hit, $c\in[0,\infty)$, iff

$$F_{T+1} \in [Y_{T+1} - c, Y_{T+1} + c].$$

Here, the hit-interval has always the length 2c, independently of Y_{T+1}. In a similar fashion it is possible to calculate combination weights defined by

$$w_{i,c} := \frac{\sum_{t=T-v+1}^{T} I_{[Y_t - c,\, Y_t + c]}(F_{it})}{\sum_{j=1}^{n}\sum_{t=T-v+1}^{T} I_{[Y_t - c,\, Y_t + c]}(F_{jt})}, \quad i = 1,\ldots,n.$$

Again, c should be selected such that there exists at least one hit in the past. It is also possible to define the hit interval as $[Y_{T+1} - c_1, Y_{T+1} + c_2]$, where $c_1, c_2 \geq 0$, $c_1 \neq c_2$.

So far we considered only combination methods which use weights depending on the number of past hits of the individual forecasts. We can also calculate weights to reach a high number of hits in the past with respect to a special minimization criterion (here MSE). Therefore, we proceed as follows:

Let $F_{1,T+1},\ldots,F_{n,T+1}$, $n \geq 2$, be unbiased forecasts for Y_{T+1} and $\mathbf{F}_{iv} := (F_{i,T-v+1},\ldots,F_{i,T})'$, $i = 1,\ldots,n$, $\mathbf{Y}_v := (Y_{T-v+1},\ldots,Y_T)'$, $v \in \{1,\ldots,T\}$, be v past observations which are used for the calculation of the weights. Further, let $\mathbf{u}_{iv} := \mathbf{Y}_v - \mathbf{F}_{iv}$, $i = 1,\ldots,n$, and $\Sigma := (\sigma_{ij})_{i,j=1,\ldots,n}$ denotes the covariance matrix of the forecast errors which we estimate in the following by $\hat{\Sigma} := (\hat{\sigma}_{ij})_{i,j=1,\ldots,n}$, where $\hat{\sigma}_{ij} := \frac{1}{v}\mathbf{u}_{iv}\mathbf{u}_{jv}'$, $i, j = 1,\ldots,n$. Then:

step 1: $\min_{\mathbf{c}} \mathbf{c}'\hat{\Sigma}\mathbf{c}$

with respect to $\mathbf{a} \leq \tilde{\mathbf{F}}_v \mathbf{c} \leq \mathbf{b}$,

where $\tilde{\mathbf{F}}_v := (\tilde{\mathbf{F}}_{1v},\ldots,\tilde{\mathbf{F}}_{nv})$, $\tilde{\mathbf{F}}_{iv} := \left(1, \mathbf{F}_{iv}'\right)'$, $i = 1,\ldots,n$, $\mathbf{c} := (c_1,\ldots,c_n)'$, $\mathbf{a} := (1, g_{T-v+1},\ldots,g_T)'$, $\mathbf{b} := (1, m_{T-v+1},\ldots,m_T)'$ and the g's and m's are the lower and upper bounds of the intervals (at the time points in the indices) given with respect to Definition 1 or 2. If the minimization problem has no solution: go to the next step.

$\vdots$

step k: Proceed as in step 1 but use $v - k + 1$ instead of v (also for the estimation of Σ).

At first we try to calculate an unbiased forecast combination with minimal variance resulting in v hits in the v most recent data points. If this is not possible we try to find a combination with $v-1$ hits in the $v-1$ most recent data points, and so on. The 1's in $\tilde{\mathbf{F}}_v$, **a** and **b** restrict the combination weights to sum up to 1. We could restrict the weights further to be non-negative. Next, we should select the bounds in the algorithm so that it stops for an "adequate" k, e.g. the number of data points considered should not be so small such that $\hat{\Sigma}$ is singular.

Another strategy is to focus in the k-th step on the v most recent data points and try to find combination weights which result in $v-k+1$ hits, but this is excluded from the following analysis.

Above, we discussed the hits-and-misses criterion based on sharp bounds between hit values and miss values. Now we will discuss approaches of hits-and-misses based on a principle well known in fuzzy theory.

Defintion 3: The hit value of a forecast F_{T+1} for a variable Y_{T+1} with respect to the functions g_1 and g_2 is given by

$$h_{c_1,c_2}(F_{T+1},Y_{T+1}) := \begin{cases} 0 & \text{if} \quad F_{T+1} \in (-\infty, Y_{T+1}-c_2) \text{ or } F_{T+1} \in (Y_{T+1}+c_2, \infty) \\ g_1(F_{T+1},Y_{T+1}) & \text{if} \quad F_{T+1} \in [Y_{T+1}-c_2, Y_{T+1}-c_1) \\ g_2(F_{T+1},Y_{T+1}) & \text{if} \quad F_{T+1} \in (Y_{T+1}+c_1, Y_{T+1}+c_2] \\ 1 & \text{if} \quad F_{T+1} \in [Y_{T+1}-c_1, Y_{T+1}+c_1] \end{cases},$$

where $c_1, c_2 \in \mathbb{R}^+$, $c_1 \le c_2$, $0 \le g_1, g_2 \le 1$ and g_1 is monotone increasing and g_2 is monotone decreasing.

In Definition 3 we can also use values depending on Y_{T+1} as in Definition 1 instead of constants c_1 and c_2. Furthermore, we can see that Definition 1 and Definition 2 are special cases of Definition 3 by letting $g_1 = 0$ and $g_2 = 0$.

Now it is possible to calculate combination weights based on Defintion 3. Again, let $F_{1,T+1},...,F_{n,T+1}, n \ge 2$, be forecasts for Y_{T+1} and F_{it}, Y_t, $i = 1,...,n$, $t = 1,...,T$ be past observations. Then we define

$$\tilde{w}_{i,c_1,c_2} := \frac{\sum_{t=T-v+1}^{T} h_{c_1,c_2}(F_{it},Y_t)}{\sum_{j=1}^{n} \sum_{t=T-v+1}^{T} h_{c_1,c_2}(F_{jt},Y_t)}, \quad i = 1,...,n,$$

where $v \in \{1,...,T\}$. The constants c_1, c_2 should be chosen so that the denominator is non-zero. Then the weight of an individual forecast is the sum of its hit values of the most recent v observations divided by the sum of hit values of all individual forecasts (of the most recent v observations).

All of the presented methods are based on the performance of the individual forecasts in the past. This is a very popular principle in combining forecasts: To use for the future what performed well in the past. If we now focus on the given expression in Definition 2 as a gain function, that is $G(u_{comb}) := I_{[-c,c]}(u_{comb})$, where u_{comb} denotes the combined forecast error, then it is possible to derive the forecast combination with maximal expected gain $E(G(u_{comb})) = P(-c \le u_{comb} \le c)$. If we additionally assume that the errors $u_{i,T+1} := Y_{T+1} - F_{i,T+1}$, $i = 1,...,n$, of the individual forecasts, are normally distributed, that is $\mathbf{u} := (u_{1,T+1},..,u_{n,T+1})' \sim N(\mathbf{0},\Sigma)$, then the "hit-optimal" unbiased forecast combination is obviously that with minimal error variance. Thus, the optimal weights are given by $\mathbf{w}_{opt} := (\mathbf{1}_n' \Sigma^{-1} \mathbf{1}_n')^{-1} \Sigma^{-1} \mathbf{1}_n$, where $\mathbf{1}_n$ denotes the $n \times 1$ vector of 1's, and we can get the weights by estimating Σ as above. These weights are identical to the MSE-optimal weights discussed in many other papers.

Finally, we compare the performance of the forecasts by a statistical test. Before combining forecasts, it is often of interest to analyse the individual forecasts. There is the demand to know if some of the forecasts are dominated by others or if the forecasts are systematically wrong. This could lead to a reduction of the number of forecasts included in the combination. Hendriksson and Merton (1981) analysed the quality of a forecast with a test based on the turning-point criterion. Their test is similar to Fisher's exact test (Cumby and Modest, 1981). We now compare two individual forecasts on the basis of Fisher's exact test where we assume that the hit probabilities are constant over time, independent of the magnitude of the variable to be forecasted. If the null hypothesis is that the two forecasts have the same quality, they should result in the same number of hits (for given observations). We can check this with e.g. Fisher's exact test. For the combination of forecasts we can proceed as follows. If an individual forecast is outperformed by another (rejection of the nullhypothesis for a given level α), we exclude it from the combination. Finally, we calculate the arithmetic mean of the remaining forecasts.

3. Application

We consider 7 institutes which forecast 6 macro economic variables each. The given time series are of length 21, where the last 10 data points are our performance points. The detailed description of the data is given in Klapper (1998). We analyse 13 combination techniques, where in each step the calculation of the weights is based on the v=10 most recent past data points.

For methods 1 and 2 we derive the weights $w_{i,\gamma}$ where $\gamma = 0.1$ and $\gamma = 0.2$. Methods 3 and 4 are based on weights $w_{i,c}$ where $c = 1$ and $c = 3$, otherwise. For methods 5 and 6 we use the algorithm presented in Section 2 $(v = 10,\ \gamma = 0.5$ and $v = 10,\ c = 2)$ where we do not restrict the weights to be non-negative since then we get in some cases no result. Methods 7 and 8 are based on Definition 3, where we consider $c_1 = 1,\ c_2 = 2$ and the functions g_1 and g_2 as follows:

method 7: $$g_1(F_{it}, Y_t) := \frac{1}{c_2 - c_1}\left(F_{it} - Y_t + c_2\right),\quad g_2(F_{it}, Y_t) := \frac{1}{c_2 - c_1}(Y_t + c_2 - F_{it})$$

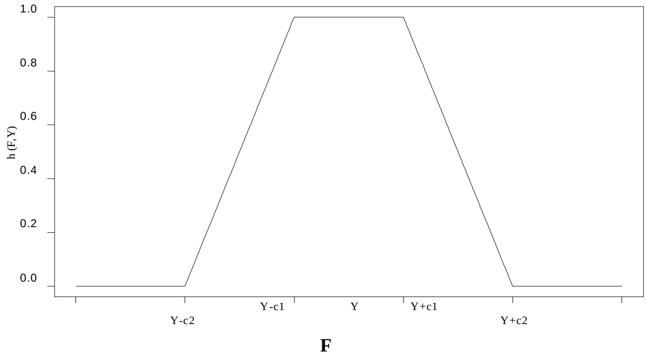
method 8: $$g_1(F_{it}, Y_t) := \frac{1}{1 - \frac{1}{\exp[(c_2 - c_1)^2]}}\left(1 - \frac{1}{\exp[(F_{it} - Y_t + c_2)^2]}\right)$$

$$g_2(F_{it}, Y_t) := \frac{1}{1 - \frac{1}{\exp[(c_2 - c_1)^2]}}\left(1 - \frac{1}{\exp[(Y_t + c_2 - F_{it})^2]}\right).$$

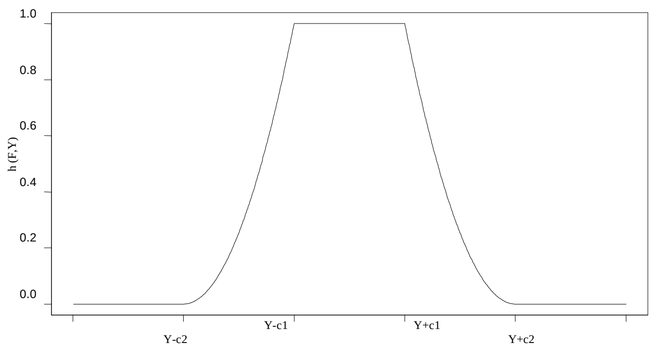
For a better illustration, the form of the function $h(\cdot,\cdot)$ related to methods No. 7 and No. 8 is shown in Figure 1.

Figure 1: Form of the function $h(\cdot,\cdot)$ in methods No. 7 and No. 8

Method No.7



Method No. 8



Method No. 9 is the MSE-optimal unbiased combination. In method No. 10 the weights are further restricted to be non-negative (in both methods the sum of weights is restricted to be 1). Method No. 11 and method No. 12 are based on the comparison of the individual forecasts by the one-sided Fisher's exact test. Therefore, we use $\alpha = 0.1$ and hit intervals given in Definition 2 with $c = 1$ and $c = 0.5$, otherwise. For the evaluation of the different techniques, we compare the RMSEs and MADs of the methods with the values of the simple average of the individual forecasts (method No. 13). The results are given in Table 1.

Table 1: Comparison of the combination techniques (relative RMSEs and *relative MADs*), M1-M13 denote the methods

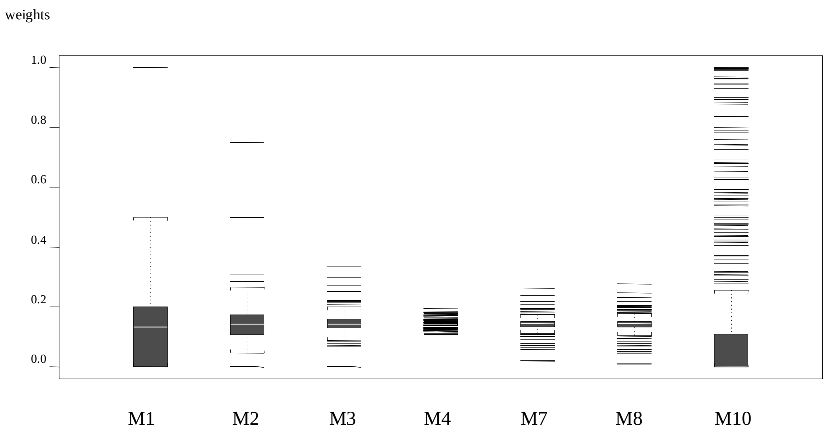
	GDP	Priv. Cons.	Publ. Cons.	Export	Import	Cons. Prices
M1	0.991 *0.989*	0.978 *9.976*	0.837 *0.820*	0.975 *0.952*	1.014 *1.012*	0.961 *0.964*
M2	1.000 *1.007*	0.998 *0.992*	0.822 *0.846*	0.974 *0.964*	1.011 *1.014*	0.993 *1.005*
M3	0.990 *0.991*	0.993 *0.984*	0.982 *0.984*	0.972 *0.947*	1.006 *1.012*	0.983 *0.987*
M4	1.000 *1.000*	0.999 *0.999*	1.000 *0.999*	0.997 *0.994*	0.998 *1.000*	1.000 *1.000*
M5	1.820 *1.453*	2.434 *2.415*	3.862 *3.438*	287.992 *112.923*	22.705 *15.374*	2.639 *2.254*
M6	1.834 *1.474*	1.750 *1.923*	2.521 *1.906*	287.998 *112.925*	22.705 *15.374*	1.806 1.684
M7	0.991 *0.986*	0.996 *0.995*	0.997 *0.998*	0.984 *0.972*	0.995 *1.001*	0.994 *0.995*
M8	0.990 *0.985*	0.995 *0.992*	0.994 *0.997*	0.982 *0.969*	1.002 *1.004*	0.994 *0.994*
M9	1.834 *1.474*	1.750 *1.923*	2.518 *1.876*	2.321 *1.956*	1.456 *1.319*	1.805 *1.683*
M10	0.916 *0.890*	1.072 *1.117*	0.930 *0.955*	0.939 *0.977*	0.935 *0.909*	1.059 *1.059*
M11	1.000 *1.000*	1.000 *1.000*	1.000 *1.000*	0.988 *0.982*	1.000 *1.000*	0.978 *0.980*
M12	0.983 *0.984*	1.000 *1.000*	1.001 *1.001*	0.981 *0.981*	0.999 *0.998*	1.000 *1.000*
M13	1.000 *1.000*	1.000 *1.000*	1.000 *1.000*	1.000 *1.000*	1.000 *1.000*	1.000 *1.000*

For the calculation of the weights related to methods 5 and 6 we use the S-Plus module *NUOPT*. We have to remark that the bad performance of these methods (for Export and Import) is a result of the higher number of iteration steps in the algorithm above ($k = 5$). The ten forecasts of the two methods for the variable Export are 170.81, −3575.82, 4.16, 9.71, 18.52, 32.27, 9.14, 5.52, 26.18, 19.54. We see that some of the values are not realistic (see Klapper, 1998) and so we would disregard them in practice. Looking at Table 1, the methods M1-M4, M7 and M8 perform a bit better than the simple average. The methods No. 1 and 2 especially for Public Consumption are of high quality. But in general the question arises, if the improvements justify the cost for the calculation of the weights, or if we should still calculate the simple average combination. The methods No. 11 and No. 12 result nearly in the same values as the simple average. The Fisher test rarely indicates a difference between the performance of the individual forecasts. Thus in most cases the average of all individual forecasts is calculated.

A disadvantage of methods No. 5, No. 6 and No. 9 is that they often result in weights larger than 1 or smaller than 0. Especially when the covariance structure is not stable over time, the quality of these forecasts decreases. Restricting the weights to be in the interval [0,1] as in No. 10 makes the combination more robust and leads to an improvement.

In the following, for the methods with weights in [0,1] we consider the box-plots of the combination weights. We have 6 time series with 7 individual forecasts each. For all series we focus on 10 performance points and thus for each of the methods above, 420 weights are calculated. The box-plots include only the methods No. 1, 2, 3, 4, 7, 8 and 10. Methods No. 5, 6 and 9 are excluded because of the large discrepancy of the weights. Methods No. 11 and 12 because they are similar to the simple average.

Figure 2: Box-plots of weights for methods No. 1, 2, 3, 4, 7, 8, 10



The variation of the weights in methods No. 1, No. 2, No. 3 and No. 10 is higher than that for the other methods. The weights of method No. 7, No. 8 and especially No. 4 are concentrated around $\frac{1}{7}$, the weight given each individual forecast in the simple average combination. This is a reason why the latter three methods perform similarly as the simple average.

4. Simulation study for the combination of two biased forecasts

In Section 3 the individual forecasts can be considered as unbiased. There are no systematical errors. We can also use the methods if the individual forecasts are biased which in general results in a biased combination. In this case the forecast with the highest hit-probability with the restriction that the weights sum up to 1 is given by the weight vector $\mathbf{w}_{\mu,\mathrm{opt}}$. If we assume that the individual forecast errors are normally distributed, that is $\mathbf{u} := (u_{1,T+1},..,u_{n,T+1})' \sim N(\mu,\Sigma)$, we get the vector $\mathbf{w}_{\mu,\mathrm{opt}}$ from the maximization problem:

$$\max_{\mathbf{w},\mathbf{w}'\mathbf{1}_n=1} P(-c \le \mathbf{w}'\mathbf{u} \le c) = \max_{\mathbf{w},\mathbf{w}'\mathbf{1}_n=1} \left(\Phi\left(\frac{c - \mathbf{w}'\mu}{\sqrt{\mathbf{w}'\Sigma\mathbf{w}}} \right) - \Phi\left(\frac{-c - \mathbf{w}'\mu}{\sqrt{\mathbf{w}'\Sigma\mathbf{w}}} \right) \right) \qquad (4.1)$$

where c defines the hit-interval. For the case where $\mu = \mathbf{0}$ we get the optimal unbiased forecast combination given in Section 2. If we are interested in an unbiased forecast combination, we have to solve

$$\max_{\mathbf{w},\mathbf{w}'\mathbf{1}_n=1,\mathbf{w}'\mu=0} P(-c \le \mathbf{w}'\mathbf{u} \le c) = \max_{\mathbf{w},\mathbf{w}'\mathbf{1}_n=1,\mathbf{w}'\mu=0} \left(\Phi\left(\frac{-c}{\sqrt{\mathbf{w}'\Sigma\mathbf{w}}} \right) - \Phi\left(\frac{c}{\sqrt{\mathbf{w}'\Sigma\mathbf{w}}} \right) \right)$$

and thus, following Wenzel (1999) we get

$$\mathbf{w}_{\mu,\mathrm{opt,unbiased}} := \frac{\left(\mathbf{1}_n{}'\Sigma^{-1}\mu\right)\Sigma^{-1}\mu - \left(\mu'\Sigma^{-1}\mu\right)\Sigma^{-1}\mathbf{1}_n}{\left(\mathbf{1}_n{}'\Sigma^{-1}\mu\right)^2 - \left(\mu'\Sigma^{-1}\mu\right)\left(\mathbf{1}_n{}'\Sigma^{-1}\mathbf{1}_n\right)} .$$

We now perform a simulation study for the combination of two biased forecasts to analyse the quality of the different hits-and-misses strategies. We examine methods No. 3, 4, 7, 8, 11, 12 and 13 (simple average). Method No. 14 is given by (4.1) where we use $c = 2$ with the further restriction that the weights are non-negative. The two individual forecasts are denoted by I1 and I2. The other methods are excluded because they are based on the unbiasedness assumption or because of their dependence on Y_{T+1}. Here, we do not restrict the forecast combination to be unbiased. Techniques dealing with this problem are described in Wenzel (1999).

We focus on three different bias vectors, $\mu^{(1)} = (1,2)'$, $\mu^{(2)} = (-1,1)'$ and $\mu^{(3)} = (-1,2)'$. In addition, we generate 10 covariance matrices and on their basis together with the bias, 100 time series of forecast errors of length 20 each. The first ten data points are fixed for the derivation of the first combination weights, so there are 10 points left for our analysis. In each step the combination weights are reestimated on the basis of the most recent 10 data points. For each case we calculate the average of the relative RMSEs (third number in the tables). We count how often the special methods have a smaller RMSE than the simple average combination (first number) and the best individual forecast (second number). The results are presented in the following tables.

Table 2: Comparison of the methods for $\mu^{(1)} = (1,2)'$

Cov.	M3	M4	M7	M8	M11	M12	M13	M14	I1	I2
1	75 33 0.881	93 32 0.892	78 30 0.884	86 31 0.867	36 20 0.958	22 20 0.977	- 21 1.000	85 45 0.8	79 - 0.823	0 - 1.741
2	86 3 0.823	99 1 0.881	95 3 0.836	98 2 0.825	45 1 0.927	8 1 0.988	- 0 1.000	99 7 0.637	100 - 0.620	0 - 1.599
3	99 0 0.689	25 0 0.998	61 0 0.977	95 0 0.924	97 0 0.760	0 0 1.000	- 0 1.000	100 0 0.681	100 - 0.681	0 - 1.329
4	85 16 0.826	100 13 0.858	95 20 0.820	99 19 0.806	41 7 0.919	21 7 0.961	- 6 1.000	96 23 0.686	94 - 0.674	0 - 1.724
5	98 0 0.788	93 0 0.975	82 0 0.935	99 0 0.902	54 1 0.897	0 0 1.000	- 0 1.000	100 0 0.66	100 - 0.66	0 - 1.362
6	83 3 0.863	100 1 0.937	99 2 0.870	100 2 0.858	42 0 0.933	0 1 1.005	- 1 1.000	99 4 0.666	99 - 0.660	0 - 1.498
7	30 38 1.026	44 39 1.001	46 47 1.012	40 43 1.010	3 41 1.006	0 42 1.006	- 43 1.000	37 30 1.028	7 - 1.208	50 - 1.038
8	33 80 1.076	58 90 0.969	51 88 0.994	57 89 0.979	6 79 1.042	6 84 1.026	- 86 1.000	53 91 0.991	12 - 1.361	2 - 1.869
9	69 25 0.937	39 31 0.999	32 30 1.014	56 28 0.984	55 13 0.974	32 27 0.984	- 31 1.000	66 29 0.951	68 - 0.938	1 - 1.321
10	40 73 1.064	42 82 0.996	42 80 1.018	43 83 1.011	5 77 1.019	3 75 1.023	- 79 1.000	35 79 1.029	16 - 1.314	5 - 1.558

In the case where both individual forecasts are positively biased, most of the hits-and-misses combinations outperform the simple average combination, the latter being of high quality for covariance matrices No. 7 and 10. Especially method No. 14 is highly reliable, because it is theoretically "hit-optimal". We observe that only in a few cases the best individual forecast is outperformed. One reason for this is that the different combination strategies in general have a higher absolute bias than the first individual forecast.

Table 3: Comparison of the methods for $\mu^{(2)} = (-1{,}1)'$

Cov.	M3	M4	M7	M8	M11	M12	M13	M14	I1	I2
1	68 95 0.908	91 96 0.865	71 93 0.932	82 95 0.875	6 83 1.035	5 88 1.013	- 89 1.000	89 99 0.813	11 - 1.393	0 - 2.538
2	80 52 0.878	97 45 0.890	89 57 0.850	95 60 0.837	23 37 0.975	4 33 0.998	- 32 1.000	92 76 0.778	68 - 0.903	0 - 1.988
3	8 100 1.322	0 100 1.000	35 100 1.09	49 100 1.013	0 99 1.294	0 100 1.037	- 100 1.000	21 100 1.047	0 - 3.208	0 - 3.346
4	83 81 0.811	99 77 0.815	82 75 0.797	99 84 0.760	16 54 0.990	7 54 1.015	- 55 1.000	97 95 0.687	45 - 1.087	0 - 2.469
5	24 97 1.133	24 100 0.993	60 98 0.994	73 100 0.959	3 97 1.111	1 82 1.244	- 100 1.000	51 100 0.973	0 - 1.817	0 - 2.499
6	56 98 1.029	68 99 0.951	69 98 0.924	88 100 0.884	1 97 1.031	1 84 1.203	- 99 1.000	79 100 0.860	1 - 1.798	0 - 2.788
7	64 48 0.962	86 42 0.949	68 45 0.948	78 48 0.939	14 34 1.007	8 35 0.997	- 35 1.000	71 51 0.913	0 - 1.596	65 - 0.958
8	16 100 1.552	46 100 0.997	22 100 1.351	39 100 1.117	0 100 1.016	1 100 1.066	- 100 1.000	57 100 0.978	0 - 4.730	0 - 5.464
9	33 100 1.134	42 100 0.982	46 99 1.052	51 100 0.997	0 99 1.064	1 99 1.044	- 100 1.000	52 100 0.980	0 - 3.234	0 - 2.575
10	17 94 1.164	33 99 1.036	20 95 1.121	19 96 1.078	4 98 1.028	1 98 1.028	- 99 1.000	28 97 1.098	0 - 1.971	1 - 1.975

At first we can see that methods No. 11 and 12 (based on the Fisher-test) in most cases are of lower quality than the simple average. The simple average combination performs best for covariance matrices No. 3 and 10. For the other matrices it is outperformed by some of the hits-and-misses techniques. Especially method No. 14 is again of high quality. Furthermore the individual forecast perform badly. The absolute bias of the combination techniques is lower than that of both individual forecasts.

Table 4: Comparison of the methods for $\mu^{(2)} = (-1,2)'$

Cov.	M3	M4	M7	M8	M11	M12	M13	M14	I1	I2
1	86	100	87	95	14	2	-	98	42	0
	85	90	84	88	56	60	58	100	-	-
	0.807	0.767	0.767	0.720	1.012	1.014	1.000	0.655	1.187	2.677
2	82	99	87	94	36	8	-	96	74	0
	67	58	62	68	23	28	26	88	-	-
	0.808	0.810	0.788	0.760	0.969	0.993	1.000	0.684	0.885	2.175
3	6	16	72	100	1	0	-	100	0	0
	16	100	100	100	42	100	100	100	-	-
	1.723	0.988	0.892	0.664	1.633	1.000	1.000	0.569	1.803	3.459
4	90	100	95	98	22	6	-	99	53	0
	88	86	90	92	50	45	47	99	-	-
	0.742	0.703	0.683	0.642	1.002	1.008	1.000	0.563	1.029	2.643
5	54	88	78	94	9	1	-	96	13	0
	93	94	97	100	69	87	87	100	-	-
	1.007	0.900	0.843	0.740	1.204	1.000	1.000	0.667	1.465	2.939
6	89	100	91	100	7	0	-	100	8	0
	98	98	98	98	90	90	92	100	-	-
	0.709	0.754	0.713	0.617	1.058	1.016	1.000	0.490	1.321	3.031
7	15	31	33	24	2	1	-	26	1	12
	66	88	85	85	81	83	87	85	-	-
	1.154	1.007	1.028	1.033	1.054	1.013	1.000	1.049	1.580	1.285
8	34	88	38	64	1	2	-	93	0	0
	96	100	100	100	99	99	100	100	-	-
	1.382	0.838	1.179	0.978	1.058	1.034	1.000	0.767	3.164	4.495
9	5	46	40	59	1	3	-	70	0	0
	70	100	100	100	69	93	100	100	-	-
	1.896	0.986	1.113	1.006	1.743	1.206	1.000	0.953	2.291	3.144
10	16	44	25	29	2	2	-	37	3	0
	90	97	92	97	95	97	97	96	-	-
	1.220	1.025	1.139	1.089	1.021	1.007	1.000	1.034	1.896	2.298

Again, the methods based on the Fisher-test cannot improve upon the simple average combination. The latter one is best for covariance matrices No. 7 and 10 and also of high quality for matrices No. 8 and 9. In general, the "hit-optimal" combination performs best. Depending on the covariance structure, the other methods are sometimes of a higher but in some cases also of lower quality than the simple average. The individual forecasts are outperformed by the different combination techniques.

Looking at the results of the simulation study, it is obvious that method No. 14 is the best. The estimation of the unknown parameters is quite good, based on the time stable covariance and bias strucure. If the structure is changing over time, the quality of this method would decrease. The methods based on the Fisher-test cannot improve the simple average combination. They often result in the simple average. Comparing methods No. 3 and 4, the combination technique using the wider hit-interval, performs a little bit better. These methods improve in many cases the simple average combination but cannot outperform it clearly. For methods No. 7 and 8, where we used the hit-functions given in Section 2, the same holds. The first individual forecast only in the case of positive bias is of a higher quality than the forecast combinations. We have to remark that only for a few time series we observed no past hits and thus we got no result.

5. Concluding remarks

The hits-and-misses criterion makes the evaluation of forecasts easy to understand for analysts. Therefore the given combination techniques are very plausible. There are a lot of variations of this criterion, especially the turning-point criterion. If we define a forecast in the right direction as a hit, then we can analyse this basing on hits-and-misses. Thus it is possible to calculate weights with respect to turning points (see Cicarelli, 1982).

The principle of defining indicator functions for the evaluation of forecasts is also analysed in the contents of interval forecasts. There, we look if a variable takes value in a forecasted interval (see Christofferson, 1998) whereas in the case of point forecasts we check if the forecast is in an interval around the value of the variable to be forecasted.

The hits-and-misses criterion makes it easy to imagine why a certain individual forecast has a bigger weight in the combined forecast than others. In the cases where we count for the forecast combination the hits in the past, the calculation of the weights is straightforward. Proceeding that way is associated with a lower cost than for many other techniques. Here, we do not need any assumptions about the error structure. In the given example of German macro economic data these methods perform a little better than the simple average combination, whereas techniques based on the covariance structure of the forecast errors are of bad quality, except in the case, where we restrict the weights to be non-negative. The Fisher test indicates that almost none of the individual forecasts is dominated by another one. This is a reason why some techniques produce nearly the same results as the simple average. We can also conclude that in such situations the simple average is a combination of high quality.

Most of the hits-and-misses combinations perform well in the simulation study. If the covariance structure is stable over time, the usage of this knowledge results in an improvement of the forecast quality.

The combination techniques based on the hits-and-misses criterion are also more robust than many other techniques. If the weights are e.g. based on the MSE or MAD of the individual forecasts, extreme outliers (extreme forecast errors) in the past data have great influence on them. On the other side, such an outlier results only in at most one more miss and hence, the weight of a special individual forecast in hits-and-misses combinations is very robust.

6. References

[1] **Armstrong, J.S.** (1985): "Long Range Forecasting", *2nd ed., Wiley, New York.*

[2] **Bates, J.M., Granger, C.W.J.** (1969): "The combination of forecasts", *Operational Research Quarterly 20*, 451-468.

[3] **Christofferson, P.F.** (1998): "Evaluating interval forecasts", *International Economic Review 39*, 841-862.

[4] **Cicarelli, J.** (1982): "A new method of evaluating the accuracy of economic forecasts", *Journal of Macroeconomics 4*, 469-475.

[5] **Cumby, R.E., Modest, D.M.** (1987): "Testing for market timing ability", *Journal of Financial Economics 19*, 169-189.

[6] **Dammers, E.** (1993): "Measurement in the ex post evaluation of forecasts", *Quality and Quantitiy 27*, 31-45.

[7] **Flores, B.E.** (1986): "Use of the sign test to supplement the percentage better statistic", *International Journal of Forecasting 2*, 477-489.

[8] **Flores, B.E**. (1989): "The utilization of the Wilcoxon test to compare forecasting methods: a note", *International Journal of Forecasting 5*, 529-535.

[9] **Hendrikson, R.D., Merton, R.C.** (1981): "On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills", *Journal of Business 54*, 513-533.

[10] **Klapper, M.** (1998): "Combining German macro economic forecasts using rank-based techniques", *Technical Report 19/1998. University of Dortmund.*

[11] **Russell, T.D., Adam, E.E.Jr.** (1987): "An empirical evaluation of alternative forecast combinations", *Management Science 33*, 1267-1276.

[12] **Schneider, C., Klapper, M., Wenzel, T.** (1999): "An evaluation of forecasting methods and forecast combination methods in goods management systems", *Technical Report 31/1999, University of Dortmund*.

[13] **Wenzel, T**. (1998): "Pitman-closeness and the linear combination of multivariate forecasts", *Technical Report 34/1998, University of Dortmund*.

[14] **Wenzel, T**. (1999): "Combination of biased forecasts: Bias correction or bias based weights?", *Technical Report 50/1999, University of Dortmund*.